ACEA Smart Water Project

Gabriel Scozzarro

19/1/2021

1.0 Introduction

The Acea Group is one of the leading Italian multiutility operators. Listed on the Italian Stock Exchange since 1999, the company manages and develops water and electricity networks and environmental services. Acea is the foremost Italian operator in the water services sector supplying 9 million inhabitants in Lazio, Tuscany, Umbria, Molise, Campania. The aim of this project is to analyze the data provided and propose one or more a prediction models to forecast water availability for each waterbody. To do that the type of questions that needs to be answered are for example:

How the different waterbodies operate? How data was collected? How climate data affects the water availability ? How was the behavior of the different waterbodies in the past?

This report will use the R language with additional packages.

2.0 Data

Each waterbody has unique characteristics and their attributes are not linked to each other. Datasets provided are completely independent from each other and the related features are also different. These variances are expected based upon the unique behavior and characteristics of each waterbody. It is fundamental to deepen the structure and the operation for each type of waterbody

They provide data for 4 different types of waterbody: aquifer, water spring, river and lake. Nine datasets were provided:

- **Auser**, Type: aquifer Description: This waterbody consists of two subsystems, called NORTH and SOUTH, where the former partly influences the behavior of the latter. Indeed, the north subsystem is a water table (or unconfined) aquifer while the south subsystem is an artesian (or confined) groundwater.The levels of the NORTH sector are represented by the values of the SAL, PAG, CoS and DIEC wells, while the levels of the SOUTH sector by the LT2 well.
- **Petrignano**, Type: aquifer Description: The wells field of the alluvial plain between Ospedalicchio di Bastia Umbra and Petrignano is fed by three underground aquifers separated by low permeability septa. The aquifer can be considered a water table groundwater and is also fed by the Chiascio river. The groundwater levels are influenced by the following parameters: rainfall, depth to groundwater, temperatures and drainage volumes, level of the Chiascio river.
- **Doganella**, Type: aquifer Description: The wells field Doganella is fed by two underground aquifers not fed by rivers or lakes but fed by meteoric infiltration. The upper aquifer is a water table with a thickness of about 30m. The lower aquifer is a semi-confined artesian aquifer with a thickness of 50m and is located inside lavas and tufa products. These aquifers are accessed through wells called Well 1, . . . , Well 9. Approximately 80% of the drainage volumes come from the artesian aquifer. The aquifer levels are influenced by the following parameters: rainfall, humidity, subsoil, temperatures and drainage volumes.

- **Luco**, Type: aquifer Description: The Luco wells field is fed by an underground aquifer. This aquifer not fed by rivers or lakes but by meteoric infiltration at the extremes of the impermeable sedimentary layers. Such aquifer is accessed through wells called Well 1, Well 3 and Well 4 and is influenced by the following parameters: rainfall, depth to groundwater, temperature and drainage volumes.

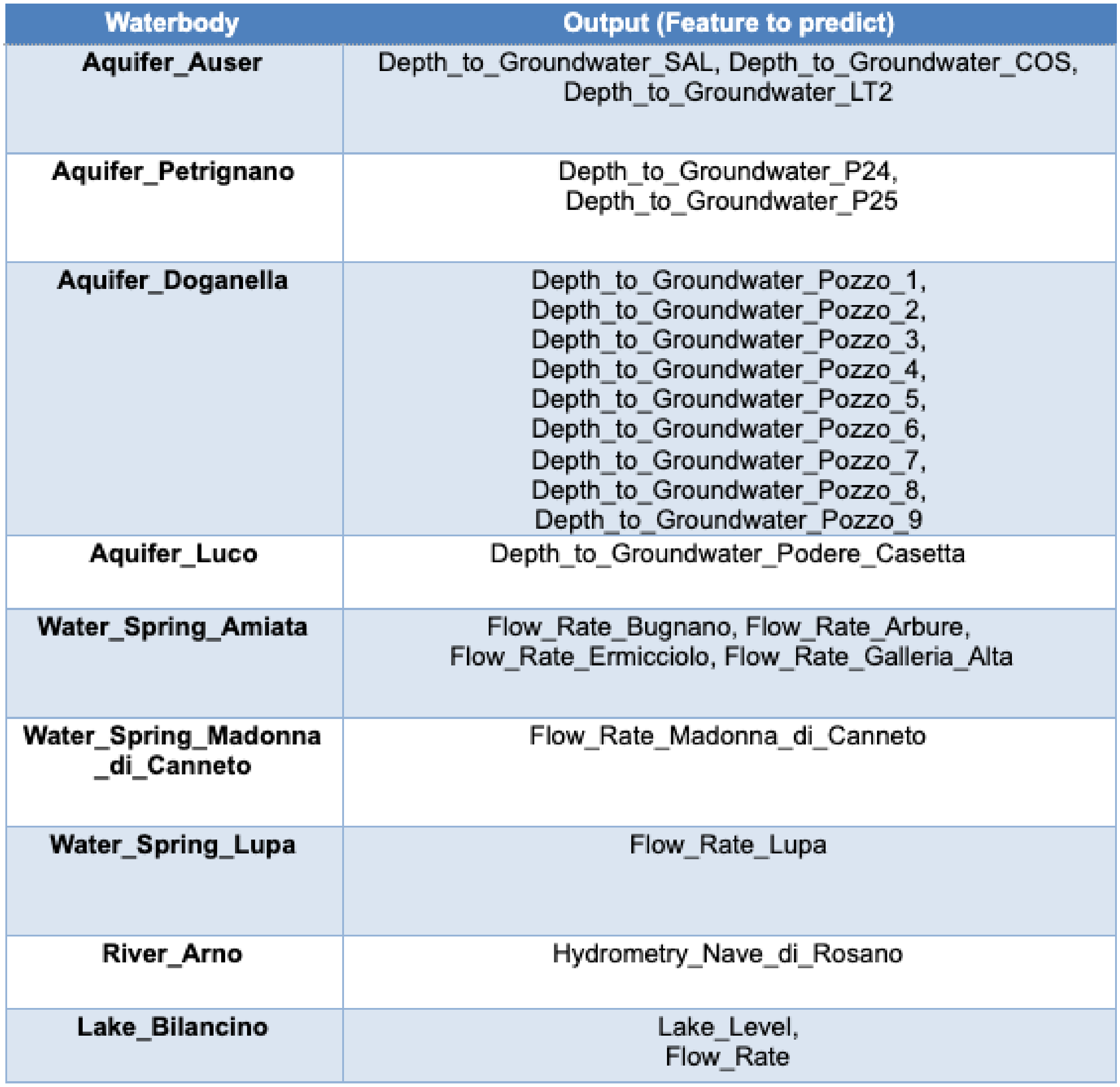
- **Amiata**, Type: water spring Description: The Amiata waterbody is composed of a volcanic aquifer not fed by rivers or lakes but fed by meteoric infiltration. This aquifer is accessed through Ermicciolo, Arbure, Bugnano and Galleria Alta water springs. The levels and volumes of the four sources are influenced by the parameters: rainfall, depth to groundwater, hydrometry, temperatures and drainage volumes.

- **Madonna di Canneto**, Type: water spring Description: The Madonna di Canneto spring is situated at an altitude of 1010m above sea level in the Canneto valley. It does not consist of an aquifer and its source is supplied by the water catchment area of the river Melfa.

- **Lupa**, Type: water spring Description: this water spring is located in the Rosciano Valley, on the left side of the Nera river. The waters emerge at an altitude of about 375 meters above sea level through a long draining tunnel that crosses, in its final section, lithotypes and essentially calcareous rocks. It provides drinking water to the city of Terni and the towns around it.

- **Arno**, Type: river Description: Arno is the second largest river in peninsular Italy and the main waterway in Tuscany and it has a relatively torrential regime, due to the nature of the surrounding soils (marl and impermeable clays). Arno results to be the main source of water supply of the metropolitan area of Florence-Prato-Pistoia. The availability of water for this waterbody is evaluated by checking the hydrometric level of the river at the section of Nave di Rosano.

- **Bilancino**, Type: lake Description: Bilancino lake is an artificial lake located in the municipality of Barberino di Mugello (about 50 km from Florence). It is used to refill the Arno river during the summer months. Indeed, during the winter months, the lake is filled up and then, during the summer months, the water of the lake is poured into the Arno river.



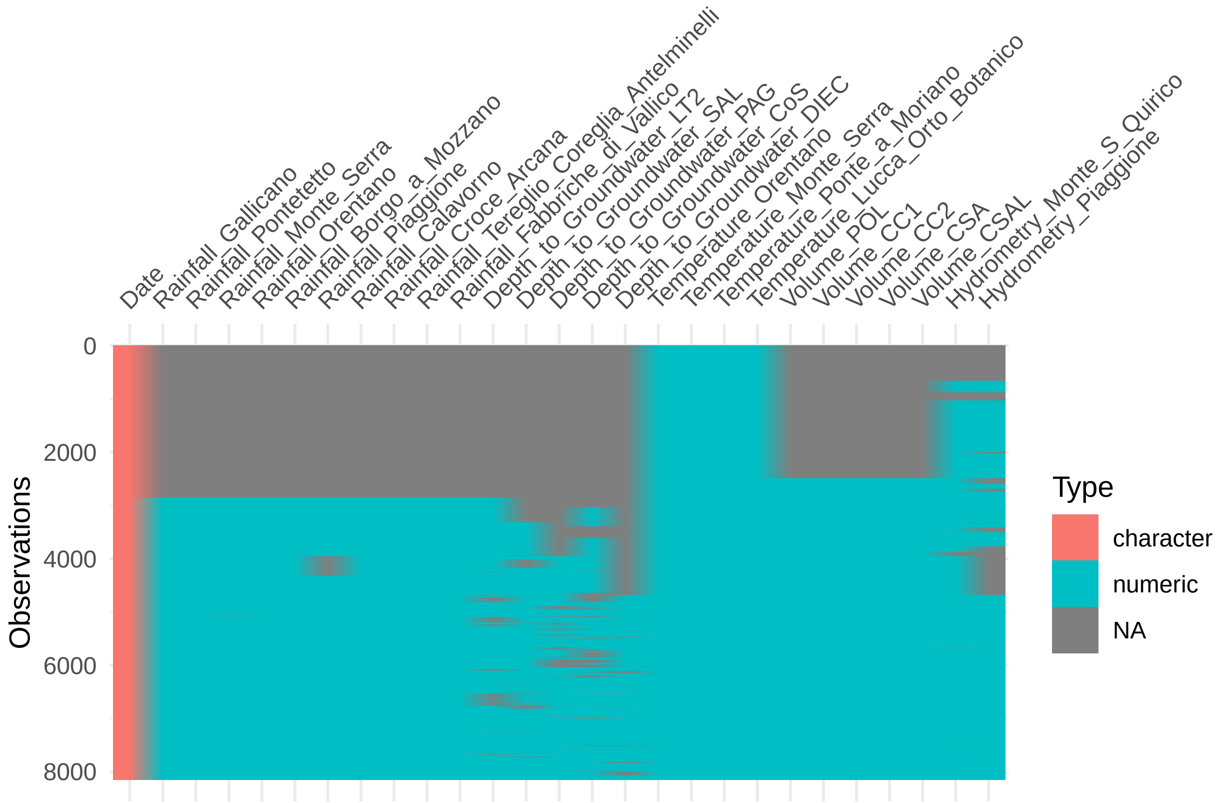
Waterbody	Output (Feature to predict)
Aquifer_Auser	Depth_to_Groundwater_SAL, Depth_to_Groundwater_COS, Depth_to_Groundwater_LT2
Aquifer_Petrignano	Depth_to_Groundwater_P24, Depth_to_Groundwater_P25
Aquifer_Doganella	Depth_to_Groundwater_Pozzo_1, Depth_to_Groundwater_Pozzo_2, Depth_to_Groundwater_Pozzo_3, Depth_to_Groundwater_Pozzo_4, Depth_to_Groundwater_Pozzo_5, Depth_to_Groundwater_Pozzo_6, Depth_to_Groundwater_Pozzo_7, Depth_to_Groundwater_Pozzo_8, Depth_to_Groundwater_Pozzo_9
Aquifer_Luco	Depth_to_Groundwater_Podere_Casetta
Water_Spring_Amiata	Flow_Rate_Bugnano, Flow_Rate_Arbure, Flow_Rate_Ermicciolo, Flow_Rate_Galleria_Alta
Water_Spring_Madonna_di_Canneto	Flow_Rate_Madonna_di_Canneto
Water_Spring_Lupa	Flow_Rate_Lupa
River_Arno	Hydrometry_Nave_di_Rosano
Lake_Bilancino	Lake_Level, Flow_Rate

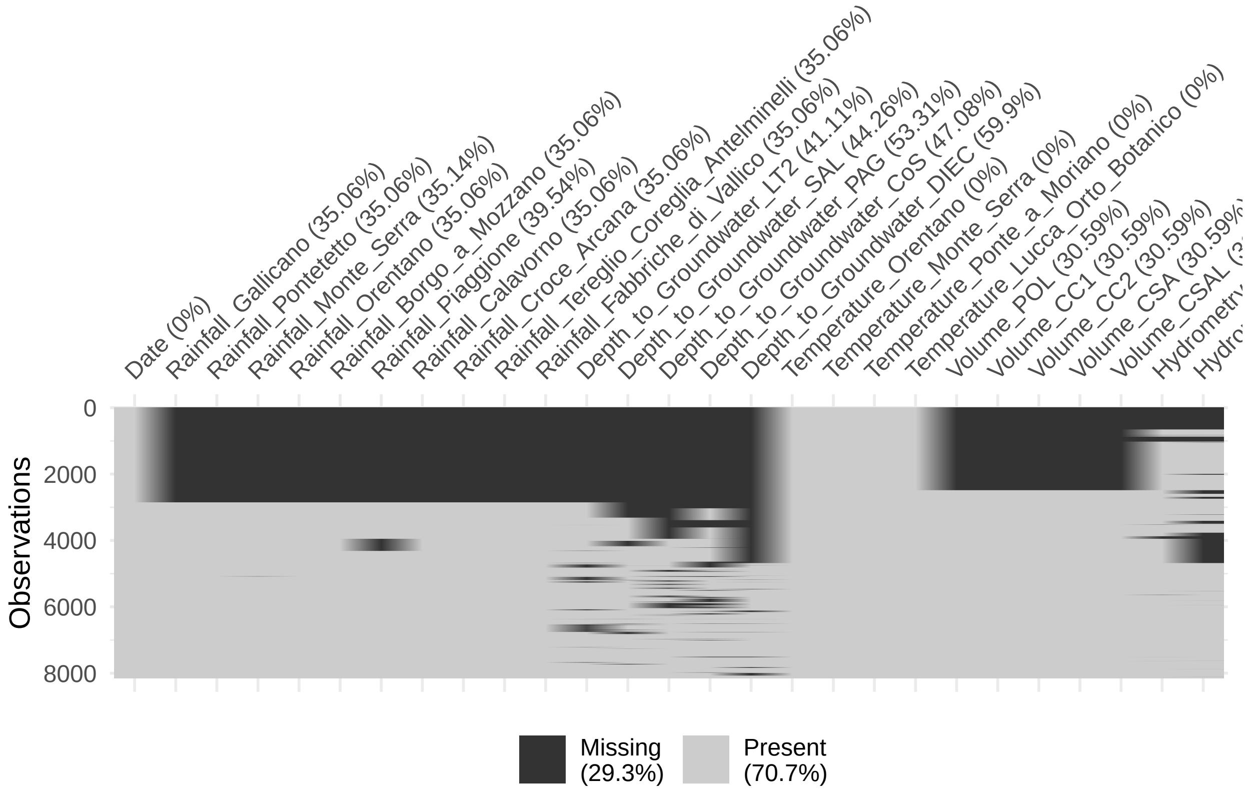
We will start working on Aquifer Auser

Table 1: Data summary table

Feature	Description	type
Date	Date	character
Rainfall_Gallicano	rainfall level	numeric
Rainfall_Pontetetto	rainfall level	numeric
Rainfall_Monte_Serra	rainfall level	numeric
Rainfall_Orentano	rainfall level	numeric
Rainfall_Borgo_a_Mozzano	rainfall level	numeric
Rainfall_Piaggione	rainfall level	numeric
Rainfall_Calavorno	rainfall level	numeric
Rainfall_Croce_Arcana	rainfall level	numeric
Rainfall_Tereglio_Coreglia_Antelminelli	rainfall level	numeric
Rainfall_Fabbriche_di_Vallico	rainfall level	numeric
Depth_to_Groundwater_LT2	level of water	numeric
Depth_to_Groundwater_SAL	level of water	numeric
Depth_to_Groundwater_PAG	level of water	numeric
Depth_to_Groundwater_CoS	level of water	numeric
Depth_to_Groundwater_DIEC	level of water	numeric
Temperature_Orentano	Local Temperature	numeric
Temperature_Monte_Serra	Local Temperature	numeric
Temperature_Ponte_a_Moriano	Local Temperature	numeric
Temperature_Lucca_Orto_Botanico	Local Temperature	numeric
Volume_POL	Volume of water used by population	numeric
Volume_CC1	Volume of water used by population	numeric
Volume_CC2	Volume of water used by population	numeric
Volume_CSA	Volume of water used by population	numeric
Volume_CSAL	Volume of water used by population	numeric
Hydrometry_Monte_S_Quirico	Local Hydrometry	numeric
Hydrometry_Piaggione	Local Hydrometry	numeric

1: Auser

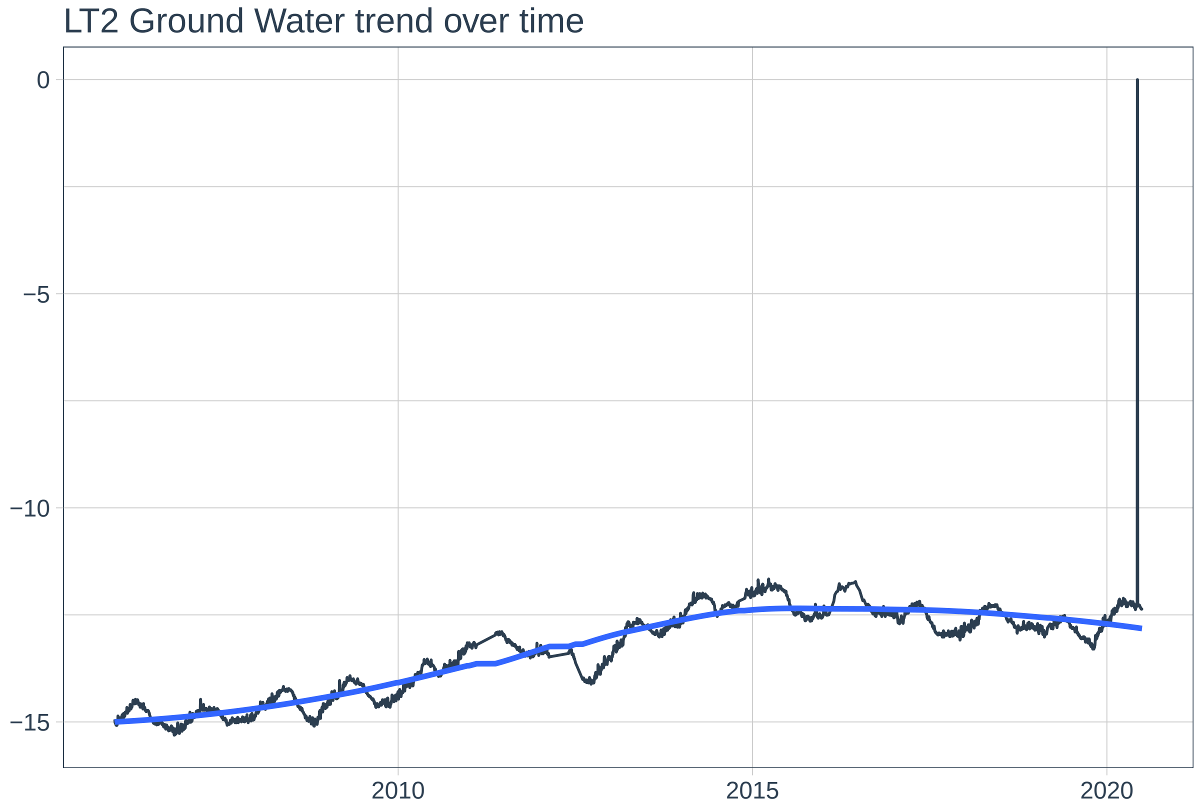

Date
Rainfall_Gallicano
Rainfall_Pontetetto
Rainfall_Monte_Serra
Rainfall_Orentano
Rainfall_Borgo_a_Mozzano
Rainfall_Piaggione
Rainfall_Calavorno
Rainfall_Croce_Arcana
Rainfall_Tereglio_Coreglia_Antelminelli
Rainfall_Fabbriche_di_Vallico
Depth_to_Groundwater_LT2
Depth_to_Groundwater_SAL
Depth_to_Groundwater_PAG
Depth_to_Groundwater_CoS
Depth_to_Groundwater_DIEC
Temperature_Orentano
Temperature_Monte_Serra
Temperature_Ponte_a_Moriano
Temperature_Lucca_Orto_Botanico
Volume_POL
Volume_CC1
Volume_CC2
Volume_CSA
Volume_CSAL
Hydrometry_Monte_S_Quirico
Hydrometry_Piaggione
0
2000
4000
6000
8000
Observations
Type
character
numeric
NA

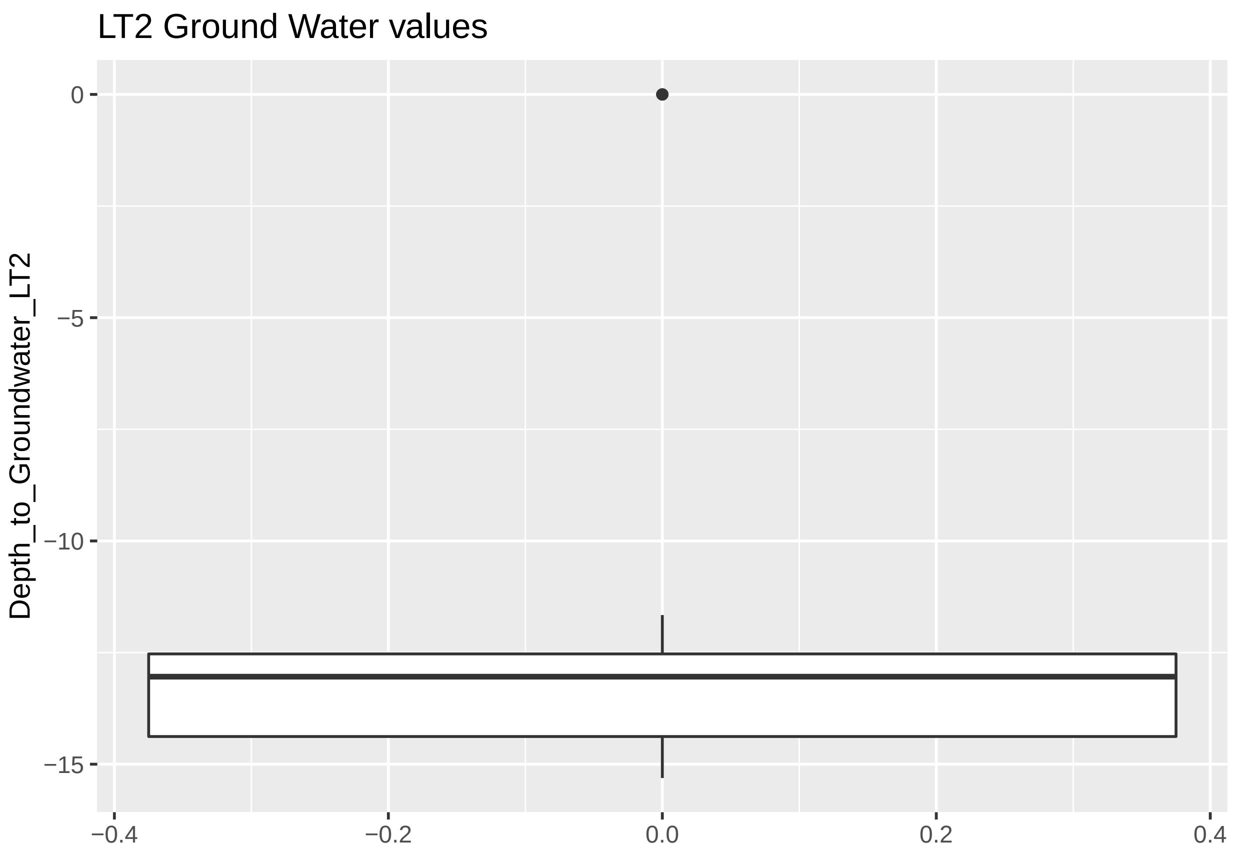

Date (0%)
Rainfall_Gallicano (35.06%)
Rainfall_Pontetetto (35.06%)
Rainfall_Monte_Serra (35.14%)
Rainfall_Orentano (35.06%)
Rainfall_Borgo_a_Mozzano (35.06%)
Rainfall_Piaggione (39.54%)
Rainfall_Calavorno (35.06%)
Rainfall_Croce_Arcana (35.06%)
Rainfall_Tereglio_Coreglia_Antelminelli (35.06%)
Rainfall_Fabbriche_di_Vallico (35.06%)
Depth_to_Groundwater_LT2 (41.11%)
Depth_to_Groundwater_SAL (44.26%)
Depth_to_Groundwater_PAG (53.31%)
Depth_to_Groundwater_CoS (47.08%)
Depth_to_Groundwater_DIEC (59.9%)
Temperature_Orentano (0%)
Temperature_Monte_Serra (0%)
Temperature_Ponte_a_Moriano (0%)
Temperature_Lucca_Orto_Botanico (0%)
Volume_POL (30.59%)
Volume_CC1 (30.59%)
Volume_CC2 (30.59%)
Volume_CSA (30.59%
Volume_CSAL (3
Hydrometry
Hydro
0
2000
4000
6000
8000
Observations
Missing (29.3%)
Present (70.7%)

The collected data in Auser aquifer dataset can be divided in mainly 4 categories describing, respectively: water depth (some of these variables will be), explanations, amount of precipitation, temperature and amount of water. We have the most complete data on temperature, there are no deficiencies across the entire database (from 03-1998). We have data on depth and precipitation from around 2006, and data on the amount of water from around 2005. You can also find individual missing data or appearing in small series here and there, in particular for depth variables.

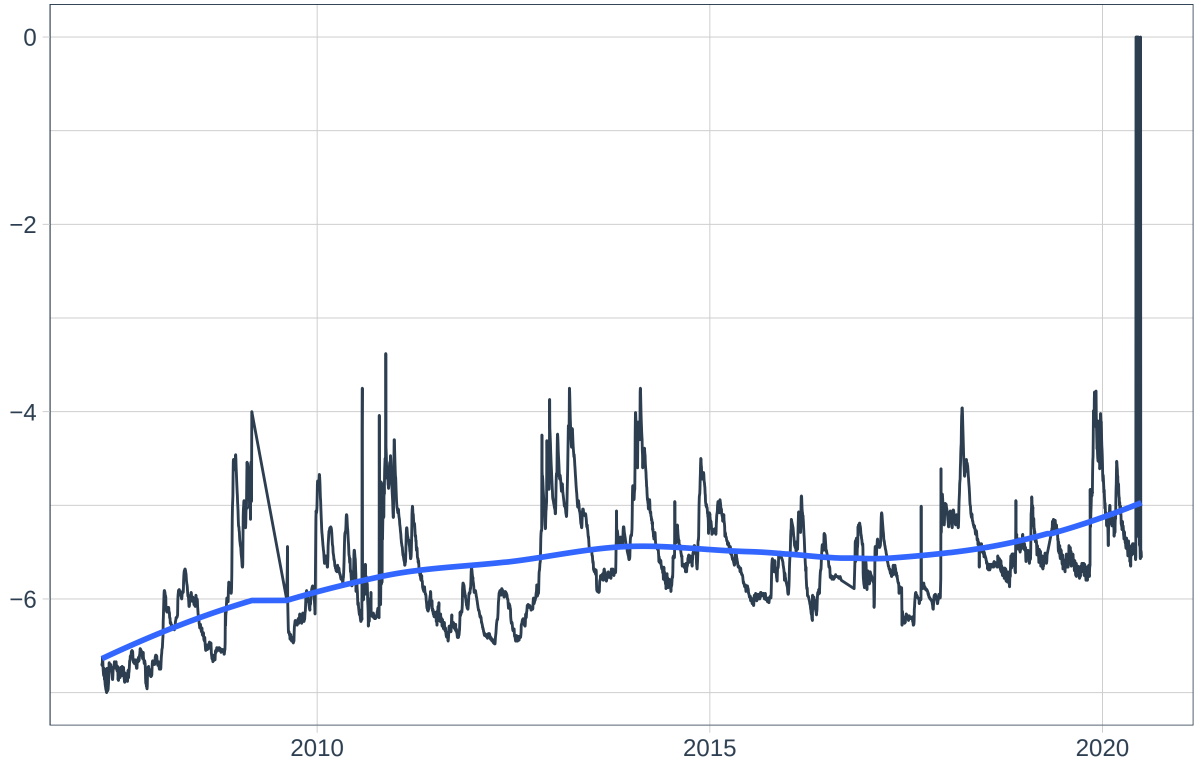
We proceed forward cleaning the dataset

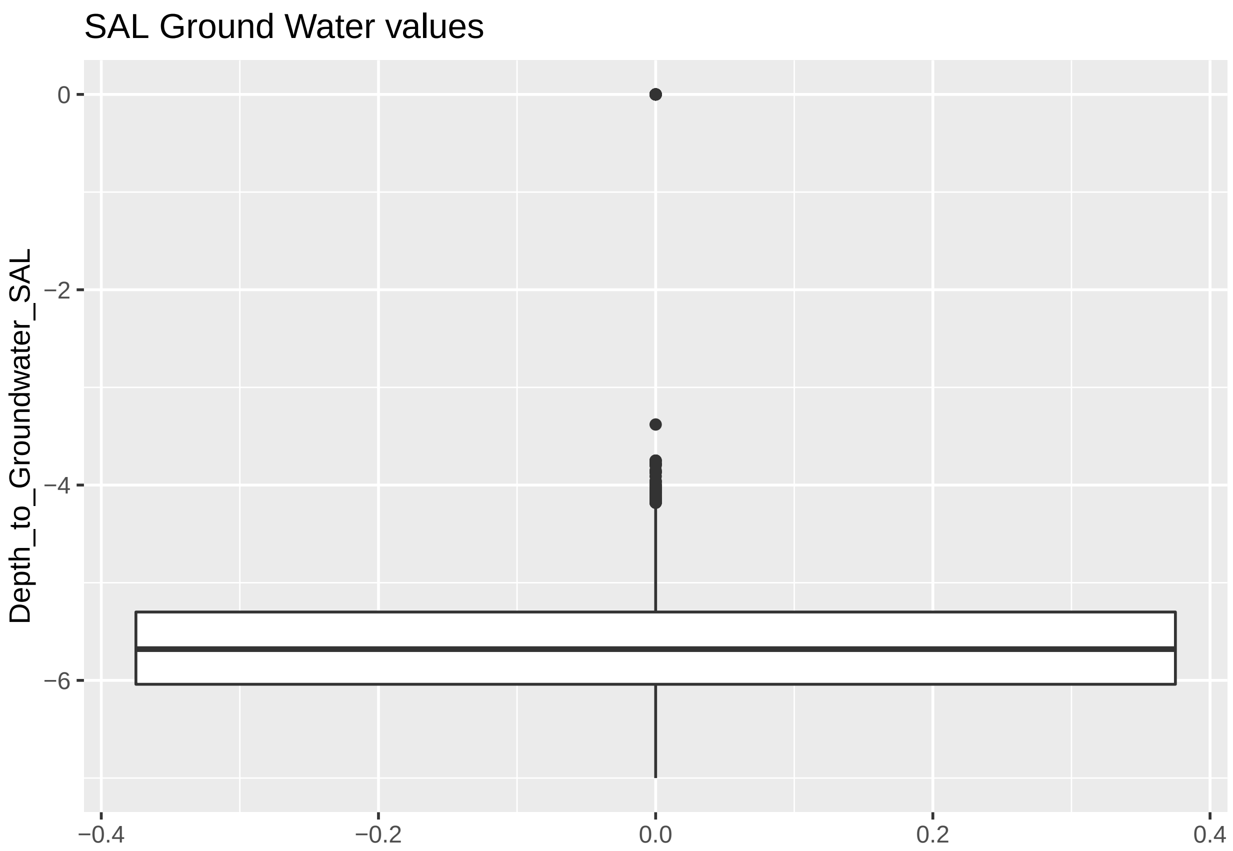
As shown before for aquifer Auser we have 3 different target value for the prediction model. Their trends over time are:




LT2 Ground Water values
Depth_to_Groundwater_LT2
0
−5
−10
−15
−0.4
−0.2
0.0
0.2
0.4

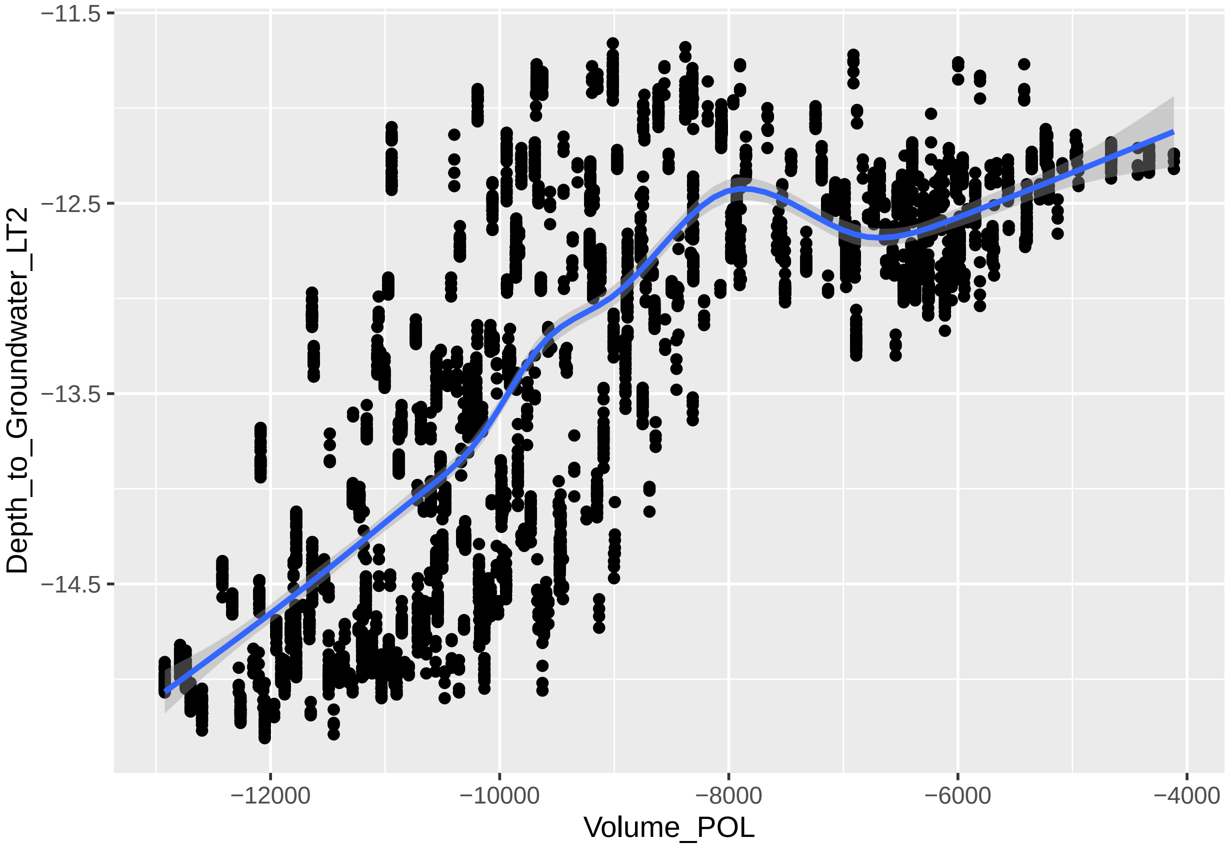
SAL Ground Water trend over time

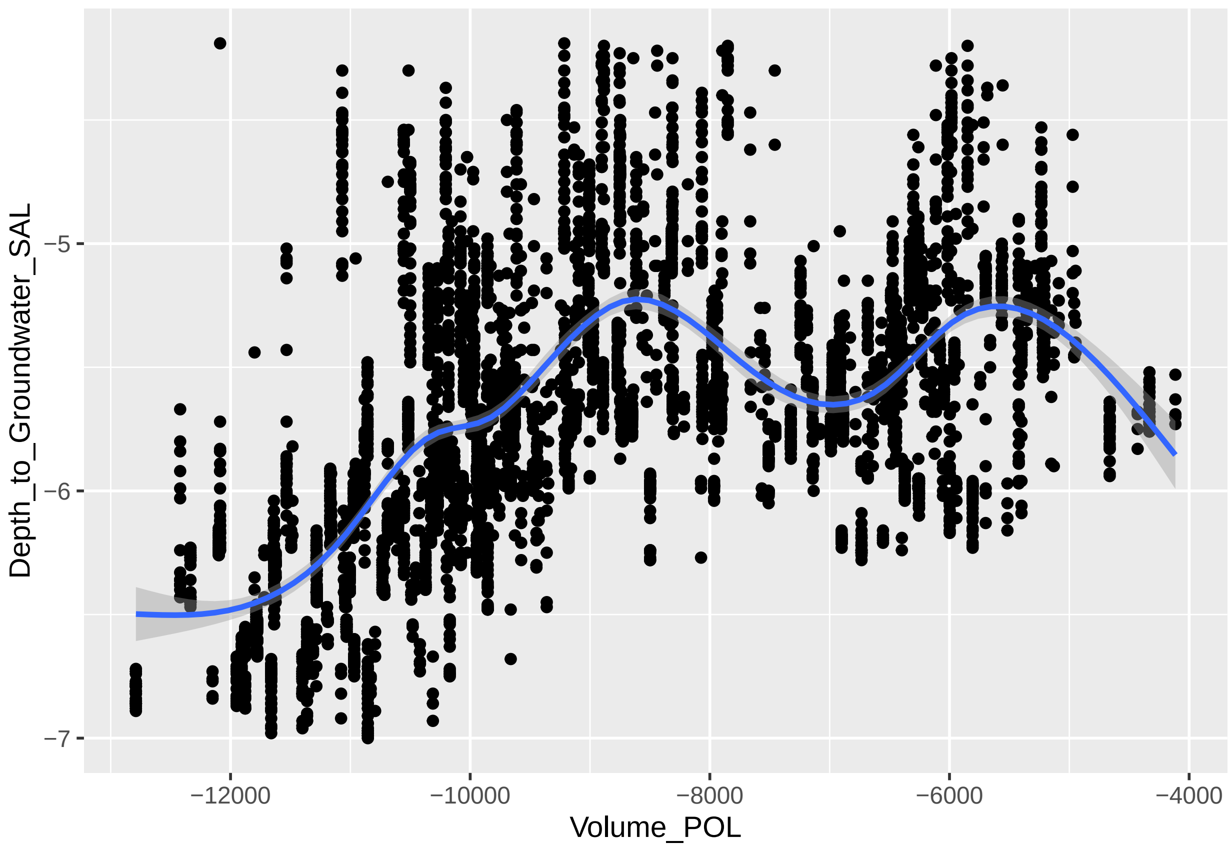




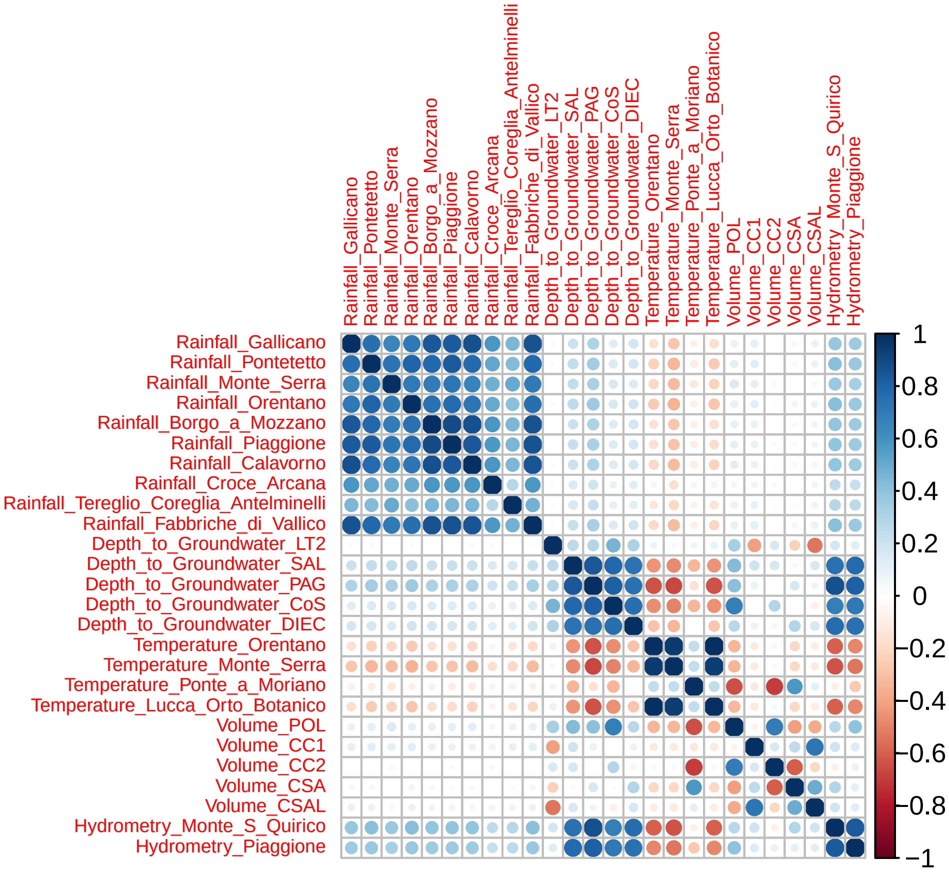
In this target values there are some serious outliers especially after 2020. This could be caused by sensors malfunction. I decide to get rid of this values that could deceive the prediction model.

Thanks to deep investigation on the operations of Auser aquifer we discover that target values are very correlated to volume of water described in Volume_POL, Volume_CC1, Volume_CC2, Volume_CSA and Volume_CSAL.


−11.5
−12.5
−13.5
−14.5
Depth_to_Groundwater_LT2
−12000
−10000
−8000
−6000
−4000
Volume_POL



Other correlation are shown as follow:



3.0 Prediction Model

Since the nature of the datasets is temporal the first prediction model approach is a Time series forecasting. The temporal horizon for the forecast is 360 days. This method will be apply on 2 target variable to asses the efficacy.

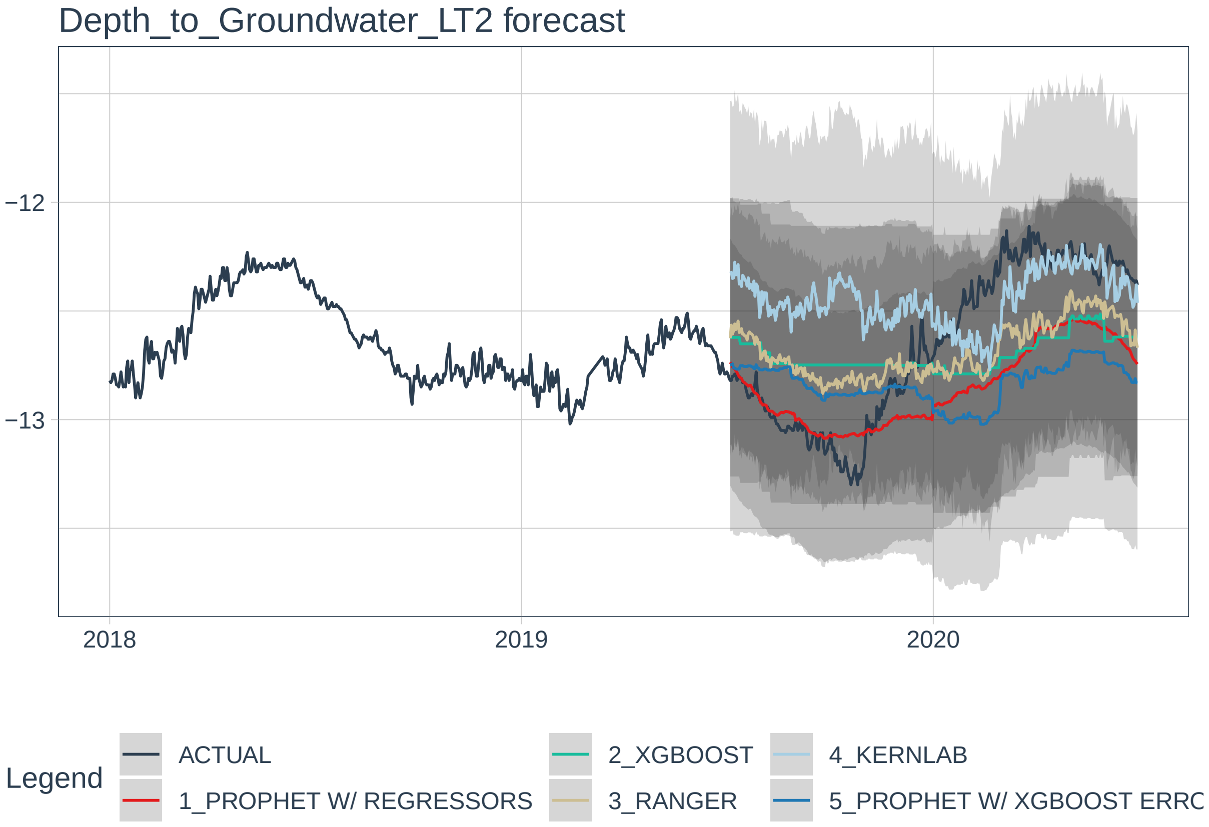
We create several models using different algorithm: Prophet, XGBoost, Random Forest, SVM and Prophet boost. Below the results obtained with this models for the prediction of Depth_to_Groundwater_LT2 an Depth_to_Groundwater_SAL.

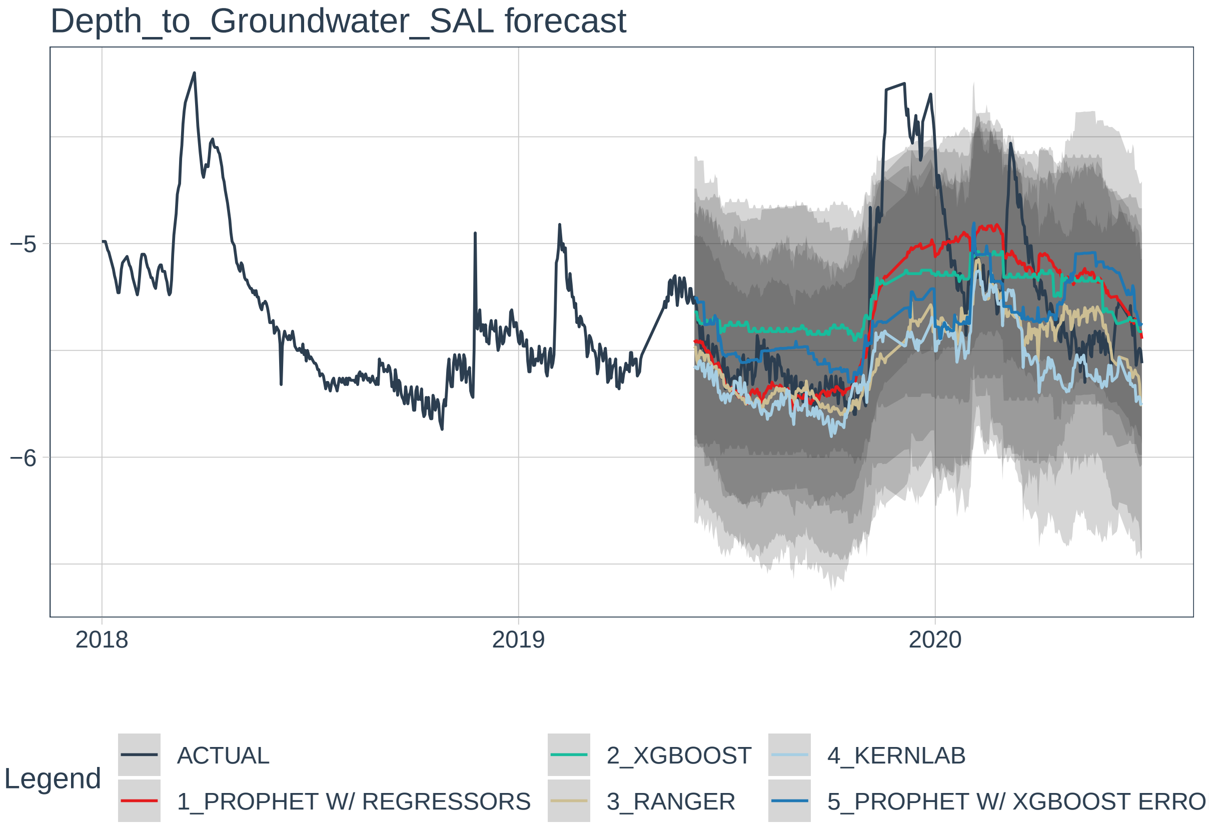
Table 2: Depth to groundwater LT2 prediction model performance

.model_id	.model_desc	.type	mae	mape	mase	smape	rmse	rsq
1	PROPHET W/ REGRESSORS	Test	0.2376035	1.909921	9.016879	1.884321	0.2900623	0.77829970
2	XGBOOST	Test	0.2994259	2.373470	11.362992	2.365170	0.3262954	0.26917262
3	RANGER	Test	0.2567269	2.032554	9.742595	2.028017	0.2782653	0.55408960
4	KERNLAB	Test	0.3186365	2.472172	12.092019	2.513862	0.4034879	0.06659873
5	PROPHET W/ XGBOOST ERRORS	Test	0.3434189	2.746437	13.032492	2.709707	0.3909277	0.05321398

Table 3: Depth to groundwater SAL prediction model performance

.model_id	.model_desc	.type	mae	mape	mase	smape	rmse	rsq
1	PROPHET W/ REGRESSORS	Test	0.1910598	3.766637	3.497729	3.722186	0.2517292	0.5295456
2	XGBOOST	Test	0.2419176	4.682693	4.428783	4.635986	0.2943180	0.5011750
3	RANGER	Test	0.2267052	4.609834	4.150289	4.362685	0.3474222	0.3611858
4	KERNLAB	Test	0.2652904	5.329404	4.856668	5.045931	0.3693593	0.5307164
5	PROPHET W/ XGBOOST ERRORS	Test	0.2479478	4.925616	4.539177	4.784514	0.3378059	0.1509607





For the prediction of Depth to groundwater LT2 as shown in the first plot and first table the 3 best models according to RMSE are PROPHET W/ REGRESSORS, XGBOOST, RANGER. The best RMSE error was 0.28, which means that the Ranger random forest model was on average wrong by 27 centimeters for the level of water indicated by the LT2 sensor.

In the same way For the prediction of Depth to groundwater LT2 as shown in the first plot and first table the 3 best models according to RMSE are PROPHET W/ REGRESSORS, XGBOOST, PROPHET W/ XGBOOST ERRORS. The best RMSE error was 0.25, which means that the Ranger random forest model was on average wrong by 25 centimeters for the level of water indicated by the SAL sensor.

To boost performance in both LT2 prediction model and SAL prediction model we try to unite and ensemble the 3 best models for each one.

Table 4: Depth to groundwater LT2 ensemble prediction model performance

.model_id	.model_desc	.type	mae	mape	mase	smape	rmse	rsq
1	ENSEMBLE (MEAN): 3 MODELS	Test	0.2574544	2.049923	9.770204	2.036093	0.2843992	0.63124795
2	PROPHET W/ REGRESSORS	Test	0.2376035	1.909921	9.016879	1.884321	0.2900623	0.77829970
3	XGBOOST	Test	0.2994259	2.373470	11.362992	2.365170	0.3262954	0.26917262
4	RANGER	Test	0.2567269	2.032554	9.742595	2.028017	0.2782653	0.55408960
5	KERNLAB	Test	0.3186365	2.472172	12.092019	2.513862	0.4034879	0.06659873
6	PROPHET W/ XGBOOST ERRORS	Test	0.3434189	2.746437	13.032492	2.709707	0.3909277	0.05321398

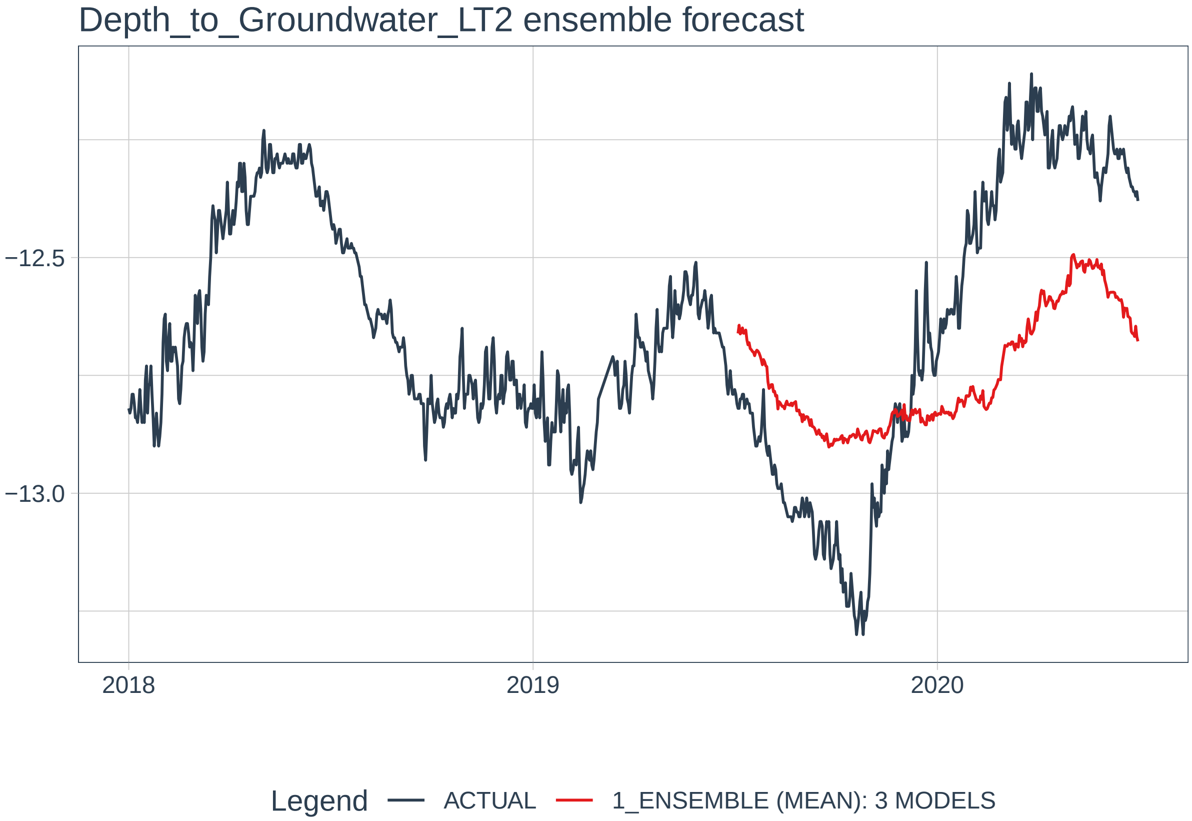
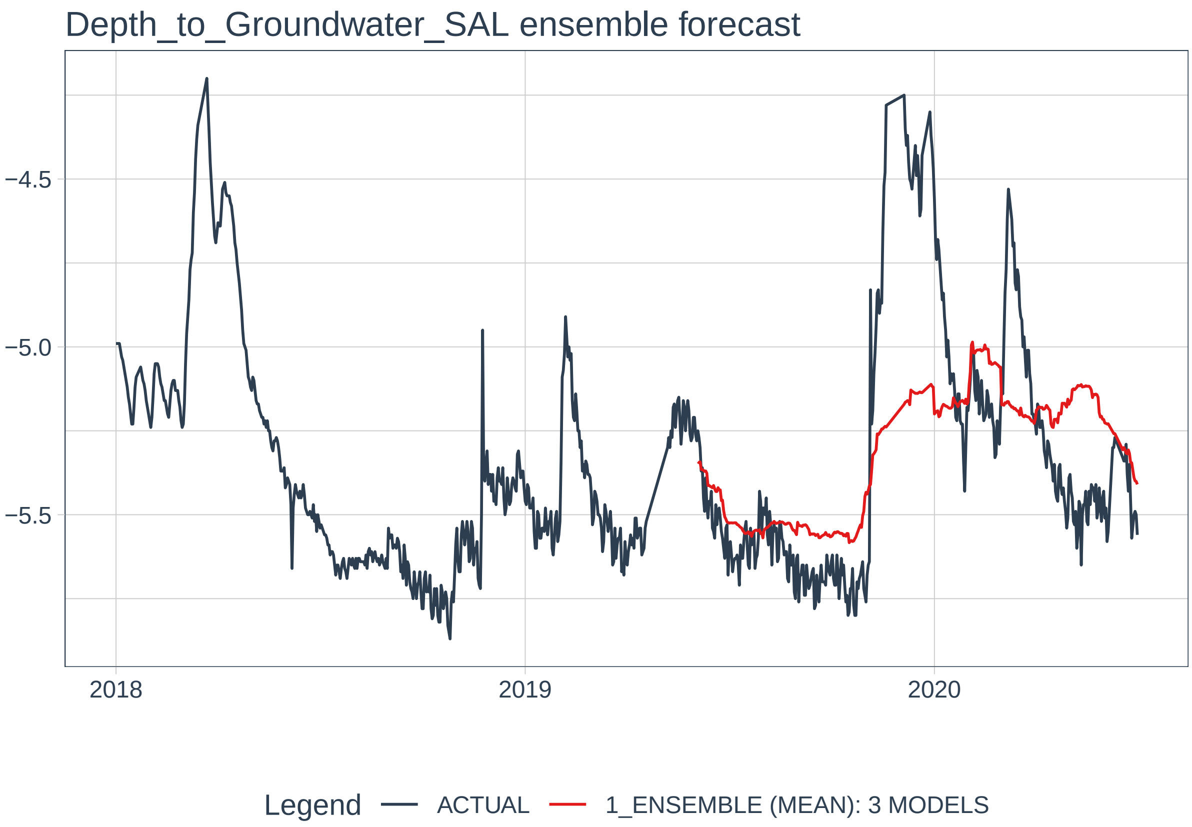


Table 5: Depth to groundwater SAL ensemble prediction model performance

.model_id	.model_desc	.type	mae	mape	mase	smape	rmse	rsq
1	ENSEMBLE (MEAN): 3 MODELS	Test	0.2048407	4.049920	3.750016	3.968943	0.2771772	0.4449918
2	PROPHET W/ REGRESSORS	Test	0.1910598	3.766637	3.497729	3.722186	0.2517292	0.5295456
3	XGBOOST	Test	0.2419176	4.682693	4.428783	4.635986	0.2943180	0.5011750
4	RANGER	Test	0.2267052	4.609834	4.150289	4.362685	0.3474222	0.3611858
5	KERNLAB	Test	0.2652904	5.329404	4.856668	5.045931	0.3693593	0.5307164
6	PROPHET W/ XGBOOST ERRORS	Test	0.2479478	4.925616	4.539177	4.784514	0.3378059	0.1509607



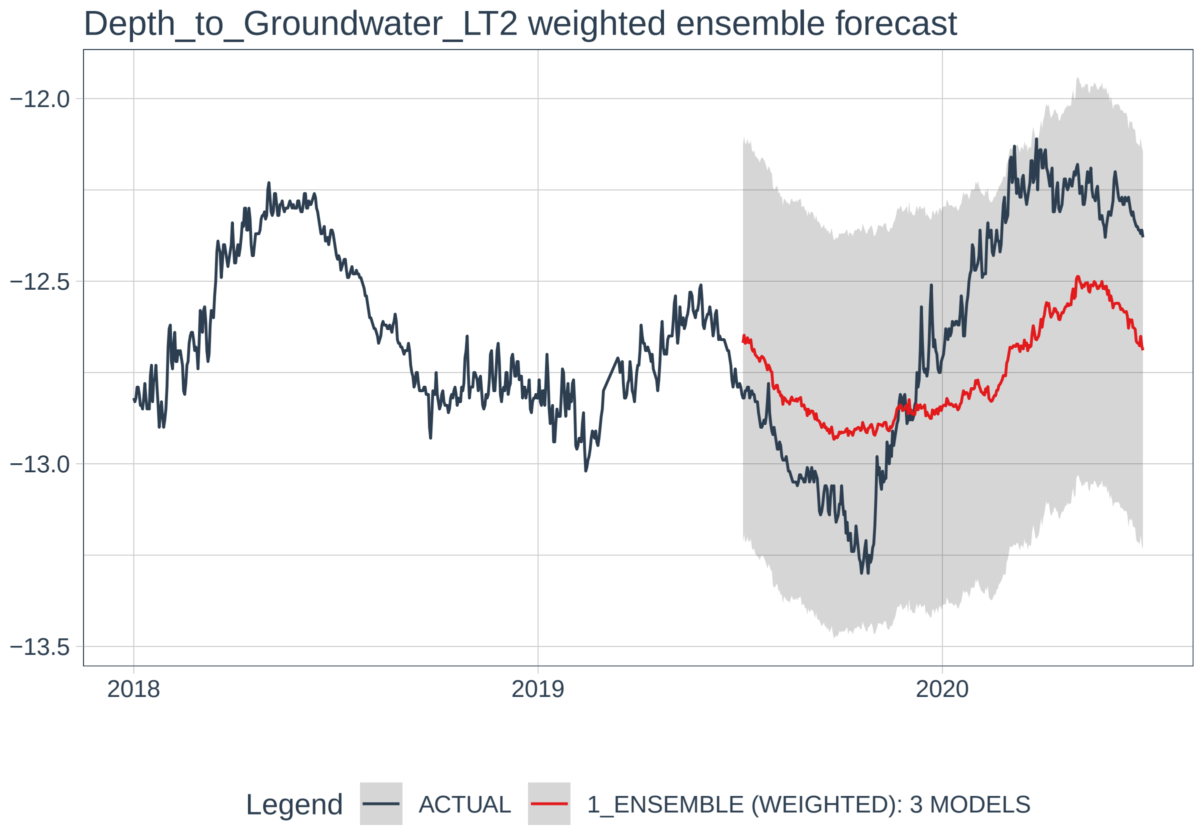
This approach is very powerful but using the mean value of each model didn't improve the RMSE. An alternative approach is weight the model in the ensemble.

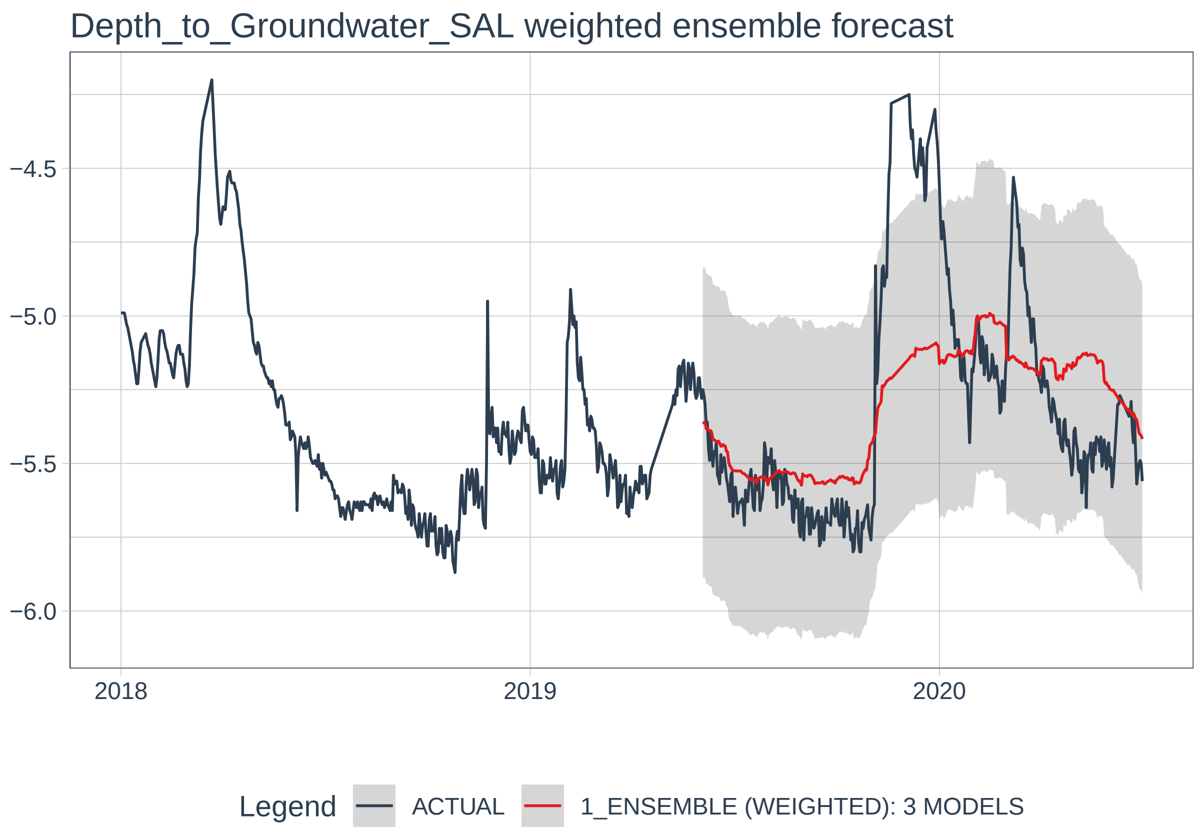
Table 6: Depth to groundwater LT2 weighted ensemble prediction model performance

.model_id	.model_desc	.type	mae	mape	mase	smape	rmse	rsq
1	ENSEMBLE (WEIGHTED): 3 MODELS	Test	0.2498206	1.991147	9.480506	1.976467	0.2777656	0.66691219
2	PROPHET W/ REGRESSORS	Test	0.2376035	1.909921	9.016879	1.884321	0.2900623	0.77829970
3	XGBOOST	Test	0.2994259	2.373470	11.362992	2.365170	0.3262954	0.26917262
4	RANGER	Test	0.2567269	2.032554	9.742595	2.028017	0.2782653	0.55408960
5	KERNLAB	Test	0.3186365	2.472172	12.092019	2.513862	0.4034879	0.06659873
6	PROPHET W/ XGBOOST ERRORS	Test	0.3434189	2.746437	13.032492	2.709707	0.3909277	0.05321398

Table 7: Depth to groundwater SAL weighted ensemble prediction model performance

.model_id	.model_desc	.type	mae	mape	mase	smape	rmse	rsq
1	ENSEMBLE (WEIGHTED): 3 MODELS	Test	0.1997499	3.942341	3.656819	3.872163	0.2679812	0.4864345
2	PROPHET W/ REGRESSORS	Test	0.1910598	3.766637	3.497729	3.722186	0.2517292	0.5295456
3	XGBOOST	Test	0.2419176	4.682693	4.428783	4.635986	0.2943180	0.5011750
4	RANGER	Test	0.2267052	4.609834	4.150289	4.362685	0.3474222	0.3611858
5	KERNLAB	Test	0.2652904	5.329404	4.856668	5.045931	0.3693593	0.5307164
6	PROPHET W/ XGBOOST ERRORS	Test	0.2479478	4.925616	4.539177	4.784514	0.3378059	0.1509607





The weighted ensemble approach is promising and worth more trials with different models even more than 3. All considered, the prophet showed great potential and can be improved using what it is called 'special date' or 'holiday' feature that need a set of dates which is correlated to a special event in the time series and so has more importance in the model.

In the next section another approach using H2O.ai that contains a number of cutting edge machine learning algorithms including Deep Learning.

Table 8: Data preparation

Depth_to_Groundwater_LT2	trend	trend_sqr	index.num	year	year.iso	half	quarter	month	month.xts	month.lbl	day	hour	minute	second	hour12	am.pm	wday	wday.xts	wday.lbl	mday	qday	yday	mweek	week	week.iso	week2	week3	week4	mday7
-14.94	1	1	1136073600	2006	2005	1	1	1	0	gennaio	1	0	0	0	0	1	1	0	domenica	1	1	1	0	1	52	1	1	1	1
-14.96	2	4	1136160000	2006	2006	1	1	1	0	gennaio	2	0	0	0	0	1	2	1	lunedì	2	2	2	1	1	1	1	1	1	1
-15.02	3	9	1136246400	2006	2006	1	1	1	0	gennaio	3	0	0	0	0	1	3	2	martedì	3	3	3	1	1	1	1	1	1	1
-15.02	4	16	1136332800	2006	2006	1	1	1	0	gennaio	4	0	0	0	0	1	4	3	mercoledì	4	4	4	1	1	1	1	1	1	1
-15.02	5	25	1136419200	2006	2006	1	1	1	0	gennaio	5	0	0	0	0	1	5	4	giovedì	5	5	5	1	1	1	1	1	1	1
-15.04	6	36	1136505600	2006	2006	1	1	1	0	gennaio	6	0	0	0	0	1	6	5	venerdì	6	6	6	1	1	1	1	1	1	1

3.1 Prediction model using H2O.ai

To use H2O.ai machine learning we will need to built a java virtual machine using R. Before that some data prepossessing was done using timetk lib augmented timeseries signature function which expands out the timestamp information column-wise into a machine learning feature set, adding columns of time series information to the original data frame. We also add a trend and trend squared features with a simple numeric index to control the upward trend and the potential non-linear trend. The data was transform as follow:

```
## Rows: 6
## Columns: 30
## $ Depth_to_Groundwater_LT2 <dbl> -14.94, -14.96, -15.02, -15.02, -15.02, -1...
## $ trend                    <int> 1, 2, 3, 4, 5, 6
## $ trend_sqr                <dbl> 1, 4, 9, 16, 25, 36
## $ index.num                <dbl> 1136073600, 1136160000, 1136246400, 113633...
## $ year                     <int> 2006, 2006, 2006, 2006, 2006, 2006
## $ year.iso                 <int> 2005, 2006, 2006, 2006, 2006, 2006
## $ half                     <int> 1, 1, 1, 1, 1, 1
## $ quarter                  <int> 1, 1, 1, 1, 1, 1
## $ month                    <int> 1, 1, 1, 1, 1, 1
## $ month.xts                <int> 0, 0, 0, 0, 0, 0
## $ month.lbl                <fct> gennaio, gennaio, gennaio, gennaio, gennai...
## $ day                      <int> 1, 2, 3, 4, 5, 6
## $ hour                     <int> 0, 0, 0, 0, 0, 0
## $ minute                   <int> 0, 0, 0, 0, 0, 0
## $ second                   <int> 0, 0, 0, 0, 0, 0
## $ hour12                   <int> 0, 0, 0, 0, 0, 0
## $ am.pm                    <int> 1, 1, 1, 1, 1, 1
## $ wday                     <int> 1, 2, 3, 4, 5, 6
## $ wday.xts                 <int> 0, 1, 2, 3, 4, 5
## $ wday.lbl                 <fct> domenica, lunedì, martedì, mercoledì, giov...
## $ mday                     <int> 1, 2, 3, 4, 5, 6
## $ qday                     <int> 1, 2, 3, 4, 5, 6
## $ yday                     <int> 1, 2, 3, 4, 5, 6
## $ mweek                    <int> 0, 1, 1, 1, 1, 1
## $ week                     <int> 1, 1, 1, 1, 1, 1
## $ week.iso                 <int> 52, 1, 1, 1, 1, 1
## $ week2                    <int> 1, 1, 1, 1, 1, 1
## $ week3                    <int> 1, 1, 1, 1, 1, 1
## $ week4                    <int> 1, 1, 1, 1, 1, 1
## $ mday7                    <int> 1, 1, 1, 1, 1, 1
```

The automachine learning function of H2O.ai was used. This function try several models and suggest the best one according to a chosen metric, which in this case it's RMSE. Above the resulting table with the model tried and the relative performances.

```
##  Connection successful!
```

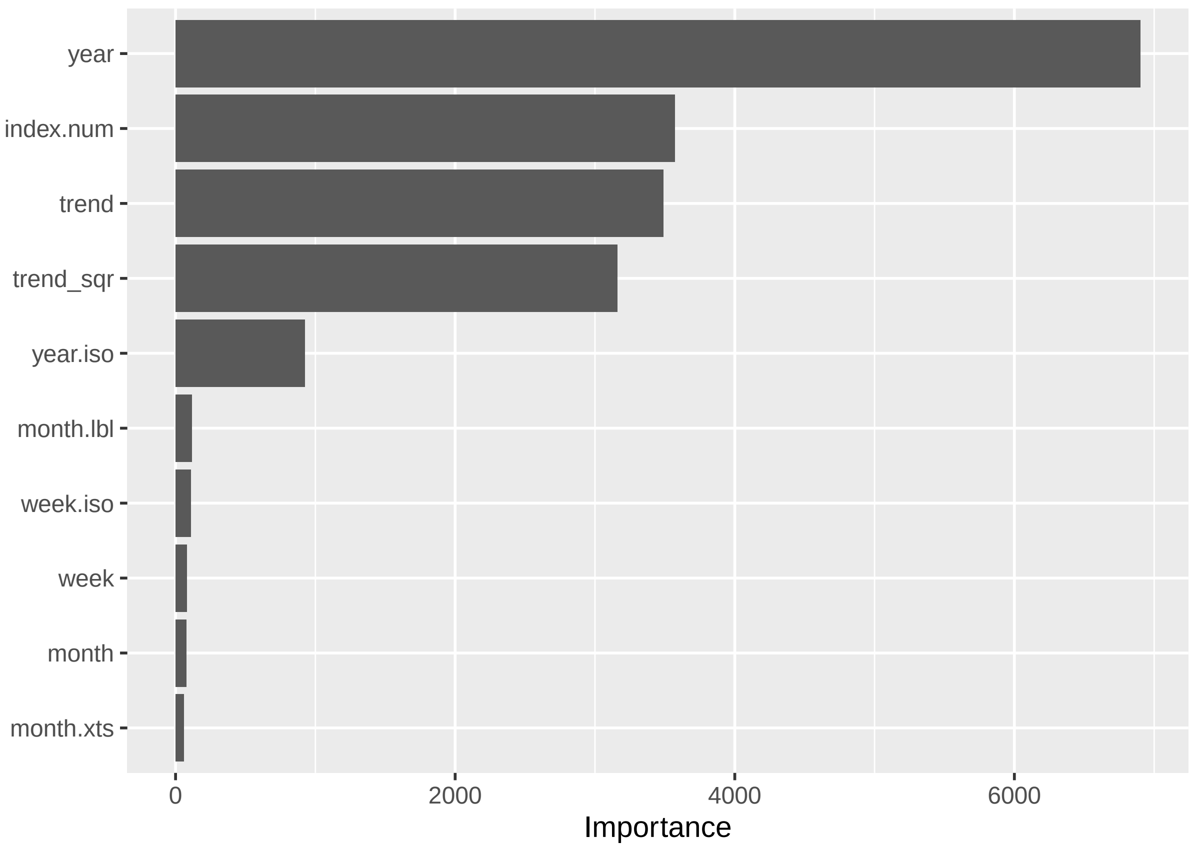
Table 9: H2o models leaderboard

model_id	mean_residual_deviance	rmse	mse	mae	rmsle
GBM_grid__1_AutoML_20210123_171342_model_82	0.01008086	0.1004035	0.01008086	0.07683704	NA
DeepLearning_grid__2_AutoML_20210123_171342_model_3	0.01314701	0.1146604	0.01314701	0.09680747	NA
GBM_grid__1_AutoML_20210123_171342_model_111	0.01412323	0.1188412	0.01412323	0.09495642	NA
GBM_grid__1_AutoML_20210123_171342_model_35	0.01467874	0.1211559	0.01467874	0.09804849	NA
DeepLearning_grid__2_AutoML_20210123_171342_model_1	0.01554757	0.1246899	0.01554757	0.10757217	NA
GBM_grid__1_AutoML_20210123_171342_model_29	0.01571347	0.1253534	0.01571347	0.10051024	NA

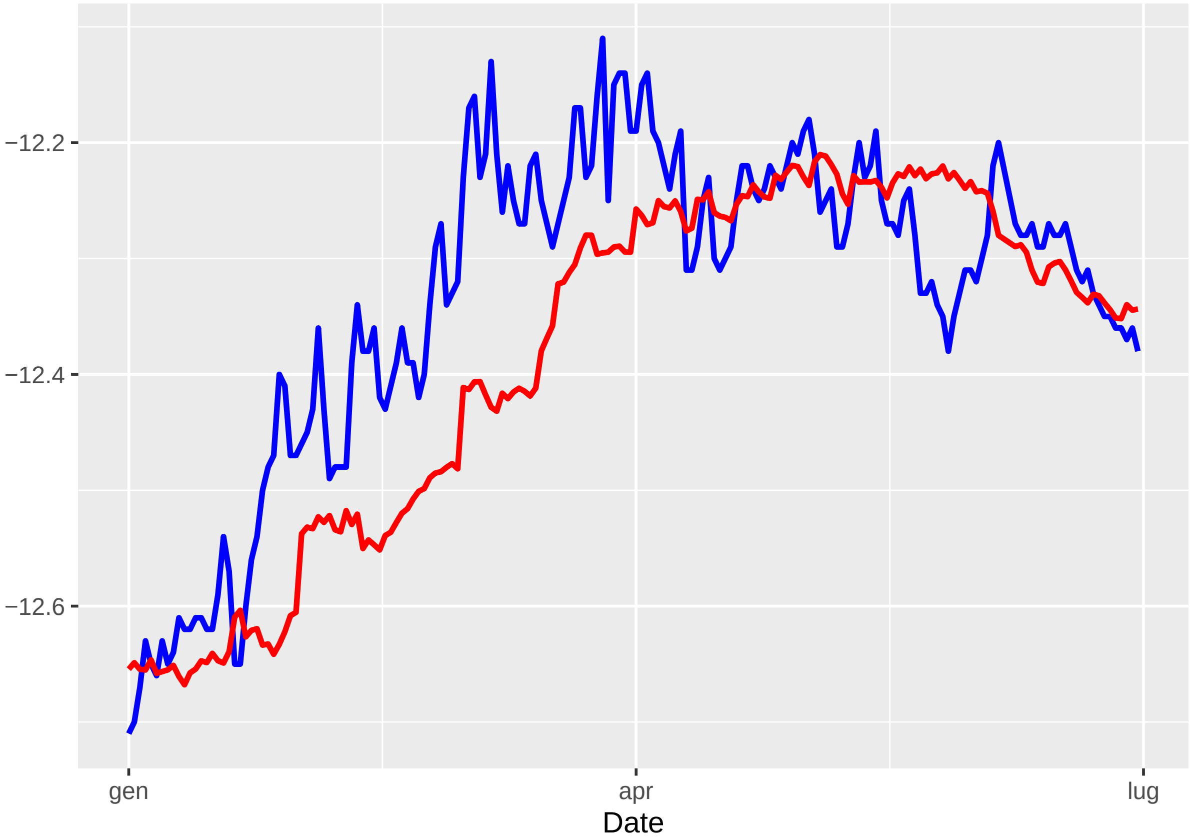
```
##
## R is connected to the H2O cluster:
##     H2O cluster uptime:         3 hours 5 minutes
##     H2O cluster timezone:       Europe/Rome
##     H2O data parsing timezone:  UTC
##     H2O cluster version:        3.32.0.1
##     H2O cluster version age:    3 months and 14 days !!!
##     H2O cluster name:           H2O_started_from_R_elekt_mql781
##     H2O cluster total nodes:    1
##     H2O cluster total memory:   5.52 GB
##     H2O cluster total cores:    12
##     H2O cluster allowed cores:  12
##     H2O cluster healthy:        TRUE
##     H2O Connection ip:          localhost
##     H2O Connection port:        54321
##     H2O Connection proxy:       NA
##     H2O Internal Security:      FALSE
##     H2O API Extensions:         Amazon S3, Algos, AutoML, Core V3, TargetEncoder, Core V4
##     R Version:                  R version 4.0.2 (2020-06-22)
```

```
##
## 17:13:42.97: User specified a validation frame with cross-validation still enabled. Please note that
## 17:13:42.97: Stopping tolerance set by the user is < 70% of the recommended default of 0.015308712907
## 17:13:42.97: AutoML: XGBoost is not available; skipping it.
```

The leader model has the follow variable importance and performance:



```
## H2ORegressionMetrics: gbm
##
## MSE:  0.01008086
## RMSE:  0.1004035
## MAE:  0.07683704
## RMSLE:  NaN
## Mean Residual Deviance :  0.01008086
```


−12.2
−12.4
−12.6
gen
apr
lug
Date